Introduction to Annif and automated subject indexing

Osma Suominen
DCMI Virtual, 21 September 2020
Osma Suominen
Information Systems Specialist, National Library of Finland

Doctoral thesis “Methods for Building Semantic Portals”
Semantic Computing Research Group, Aalto University, 2013
Supervisor Professor Eero Hyvönen

Joined the National Library in 2013
to set up the Finto.fi thesaurus and ontology service

Working on opening up bibliographic metadata as Linked Data (Fennica-LD) and automated subject indexing (Annif, Finto AI)

Open source software projects e.g.:
  Skosify - Validation and QA tool for SKOS vocabularies
  Skosmos - SKOS vocabulary publishing tool
  Annif - Tool for automated subject indexing and classification

Twitter:
@OsmaSuominen

LinkedIn:
osmasuominen

GitHub:
}@osma
Subject indexing vocabularies:

**General Finnish Ontology YSO**
(trilingual fi, sv, en, with 30,000+ concepts)

**KOKO Ontology**

...and many more

Finto.fi Where we publish thesauri, classifications, ontologies etc. for use by libraries, archives, museums, media, students...
Subject indexing

a.k.a. topic indexing, topic assignment, term assignment

~ tagging

~ multi-label classification
Extrablad till ÅBO UNDERRÅTTSELAR

Finlands oavhängighet.

Landskap som avser ingen praktiskandel om Finlands oavhängighet kan dock utgöras av det finlandssvenska rikets regering på grund av landskapets större betydelse för landets och nationens tillit.

Med till stöd för detta framförde landstinget en resolut resolution den 20 maj 1918.

Källa: Åbo Stadsarkiv

--

Extrablad till ÅBO UNDERRÅTTSELAR

Finlands oavhängighet.

Landskap som avser ingen praktiskandel om Finlands oavhängighet kan dock utgöras av det finlandssvenska rikets regering på grund av landskapets större betydelse för landets och nationens tillit.

Med till stöd för detta framförde landstinget en resolut resolution den 20 maj 1918.

Källa: Åbo Stadsarkiv
YSO, General Finnish Ontology with 30,000+ subjects
Machine learning using existing metadata

The collections of Finnish archives, libraries and museums at your fingertips.

finnaf

annif
Lexical vs. associative algorithms for subject indexing

**Lexical** approaches (e.g.: Maui)

match the terms in a document to terms in a controlled vocabulary

"Renewable resources are a part of Earth’s natural environment and the largest components of its ecosphere."

**Associative** approaches (e.g.: TF-IDF, fastText, Omikuji)

learn which concepts are correlated with which terms in documents, based on training data

Lexical approaches need comparatively little training data.

Associative approaches need a lot more training data in order to cover each subject.
Algorithms used in Annif

**Maui** (using the [Maui Server](https://www.maui-project.org) REST API)
Maui is a lexical tool for automated indexing

**TF-IDF similarity** (implemented with the [Gensim](https://radimrehurek.com/gensim/) Python library)
baseline *bag-of-words* similarity measure and vector space model

**fastText** (by Facebook Research)
uses *word embeddings* and simulates a deep *neural network* architecture

**Parabel** and **Bonsai** (implemented with the [Omikuji](https://github.com/yasuharu/omikuji) Python library)
tree-based algorithms for extreme *multi-label classification* (i.e., when the set of subjects is huge)

Implemented as Annif backends – see the [Annif wiki documentation](https://github.com/Annif/Annif/wiki) for details about each backend
Algorithms may be used **alone**, or in combinations, **ensembles**
Accessing Annif

**Command line interface**
- setup and administration
- training models
- testing and evaluating models
- bulk indexing of documents

**Web user interface**
- interactive testing of models

**REST API**
- integrating Annif services to other systems
API access example

“The quick brown fox jumped over the lazy dog.”

suggest

results=[
  {uri="http://www.yso.fi/onto/yso/p2228", score=0.2595, label="red fox"},
  {uri="http://www.yso.fi/onto/yso/p5319", score=0.2039, label="dog"},
  {uri="http://www.yso.fi/onto/yso/p8122", score=0.1946, label="laziness"},
  {uri="http://www.yso.fi/onto/yso/p25726", score=0.1285, label="brown"},
  {uri="http://www.yso.fi/onto/yso/p4760", score=0.1220, label="triple jump"}
]
Annif on GitHub

Python 3.6+ code base
Apache License 2.0

Fully unit tested (99% coverage)
PEP8 style guide compliant

Usage [documentation](https://github.com/NatLibFi/Annif) in the wiki

https://github.com/NatLibFi/Annif
pypi.org/project/annif/
quay.io/natlibfi/annif

annif 0.47.1

pip install annif

Automated subject indexing and classification tool
Apply Annif on your own data!

1. Choose subject vocabulary
2. Prepare a corpus from training data
3. Load the vocabulary and train a model
4. Suggest subjects for new documents
Demonstration of Annif

1. Load STW Thesaurus for Economics
2. Train a small model on metadata from the EconBiz portal
3. Test the model using the Web UI
Form for testing at [annif.org](http://annif.org)

Manually indexing documents for subject-based access is a very labour-intensive intellectual process. A machine could perform similar subject indexing much faster. In this series of presentations and demonstrations, we will show practical examples of automated subject indexing and discuss how such systems can be evaluated.

In the first part of this presentation, Osma Suominen will introduce the general idea of automated subject indexing using a controlled vocabulary such as a thesaurus or a classification system; and the open source automated subject indexing tool Annif, which integrates several different machine learning algorithms for text classification. By combining multiple approaches, Annif can be adapted to different settings. The tool can be used with any vocabulary; and, with suitable training data, documents in many different languages may be analysed. Annif is both a command line tool and a microservice-style API service which can be integrated with other systems. We will demonstrate how to use Annif to train a model using metadata from an existing bibliographic database and how it can then provide subject suggestions for new, unseen documents.

In the second part of the presentation, Koraljka Golub will discuss the topic of evaluating automated subject indexing systems. There are many challenges in evaluation, for example the lack of gold standards to compare against, the inherently subjective nature of subject indexing, relatively low inter-indexer consistency in typical settings, and dominating out-of-context, laboratory-like evaluation approaches.

<table>
<thead>
<tr>
<th>PROJECT (VOCABULARY AND LANGUAGE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YSO NN Ensemble English</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAX # OF SUGGESTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SUGGESTED SUBJECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>indexing</td>
</tr>
<tr>
<td>information retrieval</td>
</tr>
<tr>
<td>subject cataloging</td>
</tr>
<tr>
<td>evaluation</td>
</tr>
<tr>
<td>documentation</td>
</tr>
<tr>
<td>lists of subject headings</td>
</tr>
<tr>
<td>artificial intelligence</td>
</tr>
<tr>
<td>classification systems</td>
</tr>
<tr>
<td>automation</td>
</tr>
<tr>
<td>classification</td>
</tr>
</tbody>
</table>
Questions on the introductory part?

Next up:
Koraljka Golub: Evaluating automated subject indexing
Annemieke Romein & Sara Veldhoen: Case study on applying Annif on legal texts
Osma Suominen: Where Annif is used, evaluation results & hands-on tutorial
Annif used in production
JYX repository, University of Jyväskylä
Students upload their Master’s and doctoral theses, Annif suggests subjects*

Keywords

- information management systems [YSO]
- metadata [YSO]
- connections (technical systems) [YSO]
- content management [YSO]
- multimedia (information technology) [YSO]
- digital libraries [YSO]
- XML [YSO]
- semantic web [YSO]
- open source code [YSO]
- open data [YSO]
- user-centeredness [YSO]
- archives (memory organisations) [YSO]
- seeking [YSO]
- Works [YSO]
- cloud services [YSO]
- electronic publications [YSO]

Your own keywords

Comma separated list

keyword 1, keyword 2

Implemented using DSpace & GLAMpipe by Ari Häyrinen

*from YSO = General Finnish Ontology
Osuva repository, University of Vaasa

Same idea as JYX: students upload their theses, Annif suggests subjects

Pilot started 2.3.2020, implementation by Anis Moubarik.
Kirjavälitys Oy - logistics company serving bookstores and libraries

Publishers

information about new titles

Kirjavälitys

correction and curation

descriptive text

subject suggestions

Bookshops and online stores

Melinda inc. Fennica

Libraries
Finto AI - automated subject indexing tool and API service

Launched in May 2020

In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and animals. Leading AI textbooks define the field as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals [1]. Colloquially, the term "artificial intelligence" is often used to describe machines (or components of machines) that execute tasks that would traditionally require human intelligence, such as visual perception, speech recognition, decision making, and control of devices.

As machines become increasingly capable, terms like "superintelligent AI" or "artificial general intelligence" are often removed from the definition. For example, Martin Walser's "The Turing Test" says "AI is whatever a machine is capable of doing until it wins a competition in Tegner's Theorem says "AI is whatever a machine is capable of doing which is not predictable by humans" [3] claims having become a routine technology. Modern machine capabilities generally classified as AI include successfully understanding human speech, competing at the highest level in strategic games such as chess and Go, autonomously operating cars, intelligent routing in content delivery networks, and military simulations.

ai.finto.fi
Evaluating algorithms used in Annif
Evaluation approaches (Golub et al. 2016), emphasis mine

1. Evaluating indexing quality directly through assessment by an evaluator or by comparison with a gold standard.

2. Evaluating indexing quality directly in the context of an indexing workflow.


Assessment by evaluators

At a workshop in 2019, 48 evaluators evaluated subjects for 50 documents. Subjects were given by either human indexers or four different algorithms.

The best ensemble algorithm (red bars) was not quite on the level of human indexers in quality scores (left), and significantly more of its suggestions were rejected (right).

Annif-Leiki Comparison at Finnish Broadcasting Company Yle

- Annif vs Leiki (commercial service) tagging compared by 28 human evaluators at Yle
- About 100 Finnish and Swedish articles and their tags
  - business, science, culture, sport

Comparison: Overall Results / Finnish

Annif performed slightly better than Leiki
= more essential + ok, less not relevant + wrong tags

Culture: The only subject area where Leiki performed slightly better: more ok, less wrong tags
Annif performed better than Leiki in all subject areas = more essential + ok, less not relevant + wrong tags

Differences bigger than in Finnish

Biggest differences in business and science

Reasons?

Comparison: Overall Results / Swedish

Essential + ok (% of all tags)
TOTAL - business - sport - culture - science

Not relevant + wrong (% of all tags)
TOTAL - business - sport - culture - science
Precision, recall and F1 score

- **Precision**: fraction of the correct subjects among the subjects suggested
  “How many of the suggested subjects are actually correct?”

- **Recall**: fraction of all correct subjects that were actually suggested
  “How many of those subjects that should be suggested have actually been suggested?”

- The **F1 score** is the [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean) between precision and recall
  (i.e., a way of combining precision and recall values into one similarity score).
Comparison to “gold standard”

F1@5 scores for different test corpora and Annif API/model versions
Evaluating in the context of an indexing workflow

**JYX repository, University of Jyväskylä:**
F1 similarity between Annif suggestions and the subjects
a) chosen by the student (blue)
b) confirmed by the JYX librarian (red)

Lessons from evaluation

● The different evaluation approaches are complementary. Not a good idea to look at just a single measure.

● Improved quality of automated subject indexing over time
  ○ better training and evaluation data
  ○ better algorithms: Omikuji, neural network ensemble

● Continuous process: it never stops…
Hands-on tutorial
for those who want to learn to use Annif themselves

all materials freely available on GitHub & YouTube
Understand what Annif is
Study the website [annif.org](http://annif.org), watch a presentation about it, or read the LIBER Quarterly [paper](http://example.org).

Complete this hands-on tutorial
Watch the videos, install Annif, and complete the exercises as far as you can, on your own time.

Join an online session (optional)
In the online sessions, you can ask questions, get help and discuss what you’ve learned. Registration required.
Videos

- Introduction to the hands-on tutorial
- Install Annif: usage in VirtualBox
- Install Annif: Docker image
- Install Annif: Linux native install
- Data sets for this tutorial
- TFIDF project
- A little bit about algorithms
- Web UI
- Metrics & evaluation
- Install Maui Server: Doxygen install
- Install Maui Server: Linux native install
- 05. Maui
- Ensembles
- Closing tutorial
Core and optional exercises
Annif-tutorial [GitHub repository](https://github.com/NatLibFi/Annif-tutorial) is the main resource for the hands-on tutorial.
Online help sessions

Only register if you have watched the videos and tried to complete the exercises!

- **Friday 9 October 2020**, 07:00-09:30 UTC  
  Registration starts on Friday 25 September

- **Wednesday 21 October 2020**, 15:00-17:30 UTC  
  Registration starts on Wednesday 7 October

For more information, see the [Annif-tutorial GitHub repository](https://github.com/Annif-tutorial)
Automated Subject Indexing IG

Name: Automated Subject Indexing IG
Type:
Status: active

Charter: This group will focus on automated and semi-automated solutions for subject-based information organization of digital collections, i.e., the extraction of semantic features from textual data with methods from areas such as natural language processing and machine learning, and ways of integrating those solutions into productive subject indexing systems.

Moderator/Chair: Koraljka Golub  
Anna Kasprzik  
Osma Suominen

Established: 2019-04-04

The activities of the group include:

- exchanging ideas on possible research approaches, workflows, algorithms and best practices for automated subject indexing
- collecting and providing information on existing tools, corpora and other resources that are openly available
- working on reusable and interoperable open source tools and coordinating the various development efforts of do-it-yourself solutions within the community
- issuing recommendations for the standardization of APIs and file formats for data exchange, and for more detailed metadata schemata in order to support and document automated methods in subject indexing.

Anyone, not just DCMI members, can join the group. Please subscribe to the [autosubject-ig mailing list](https://www.dublincore.org/groups/automated_subject_indexing_ig/) if you want to take part in activities of the group.

https://www.dublincore.org/groups/automated_subject_indexing_ig/
Thank you!

Juho Inkinen  
Mona Lehtinen  
Osma Suominen

annif.org

These slides: https://tinyurl.com/annif-dcmi2020